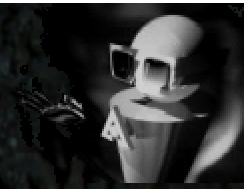
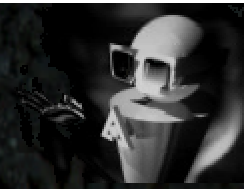


Which Node for Your Cluster?

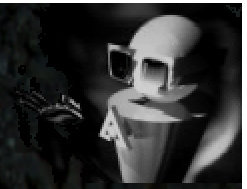
Rick Stevens
Argonne National Laboratory
University of Chicago
stevens@mcs.anl.gov

Outline of Talk



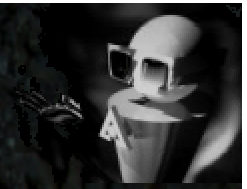
- Cluster Projects at Argonne/Overall Structure and Organization
- Interconnect Options
- Node Issues
  - Floating Point Performance
  - Memory Bandwidth
  - Cache Performance and Cache Sizes
  - Motherboards and Chip Sets
- Node Options
  - Intel Pentium III
  - AMD Athlon
  - Compaq Alpha
  - Intel IA-64
  - IBM PowerPC

Cluster Projects at Argonne



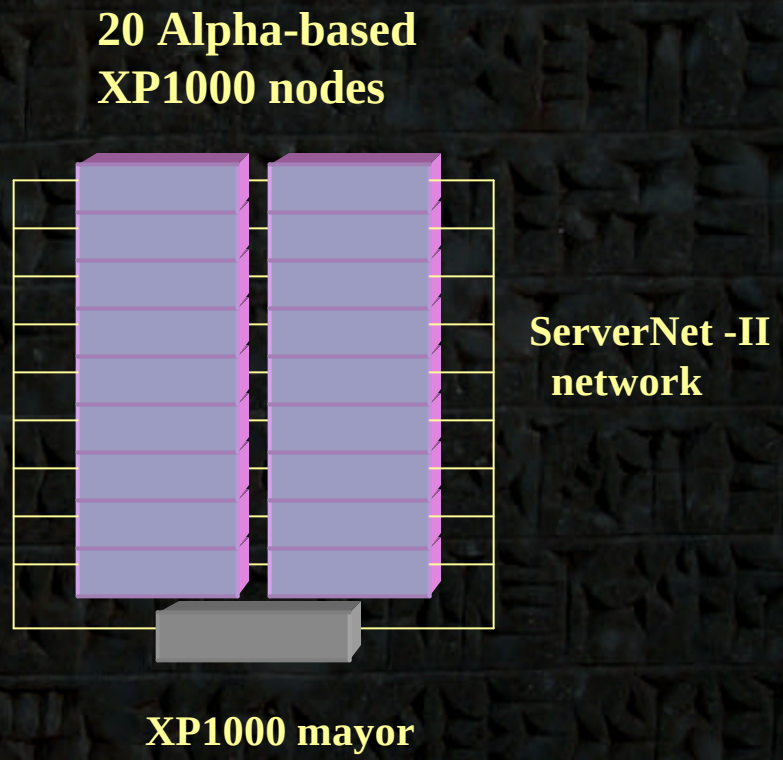
- Windows NT SuperCluster (1996-1998) (32 heterogeneous nodes)
- Windows NT MPI porting project (8 nodes)
  - Microsoft supported (PIIs,SMP, 256 MB, 8 GB disk, Giganet)
- Collage and ActiveMural (22 nodes)
  - Linux Cluster to Drive (PIIs, 256 MB, 4 GB disk, display Adapters)
- Chiba City Test Clusters (4-16 nodes)
  - various development environments
- Chiba City Main System (300+ nodes)
  - 256 core nodes (dual PIIIs, 512MB RAM, 9 GBdisk, Myrinet, Gig-E)
  - 32 visualization nodes (PIIIs, 512MB RAM, 9 GBdisk, Myrinet, Matrox G200s)
  - 8 storage nodes (Xeon, 300 GB Disk/node)
  - 18 management nodes (PIIs, multiple networks, 18 GB disk)
- Distributed Storage Server Prototype (20 nodes) 4 TB aggregate
  - PIII, 512 MB RAM, 200 GB disk, Fast and Gig-E
- Alpha Cluster for Computational Biology
  - 18 XP1000’s (633MHz, Ev6), 512 MB RAM, 10 GB disk , ServerNetII, Management Node

Argonne Alpha Cluster (Huxley)

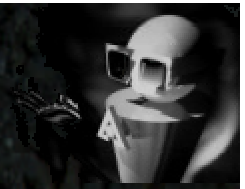


- 20 Compaq XP1000 nodes
  - 512M RAM, 9GB Disk
- 1 XP1000 mayor
- ServerNet - II network
- Switched ethernet and serial connections for management
- Alpha Linux-based

- Molecular Modeling
- Digital Cell Project
- Structure Determination
- Biological CAD



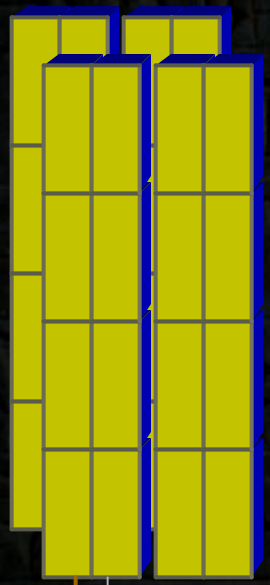
Chiba City
The Argonne Scalable Cluster

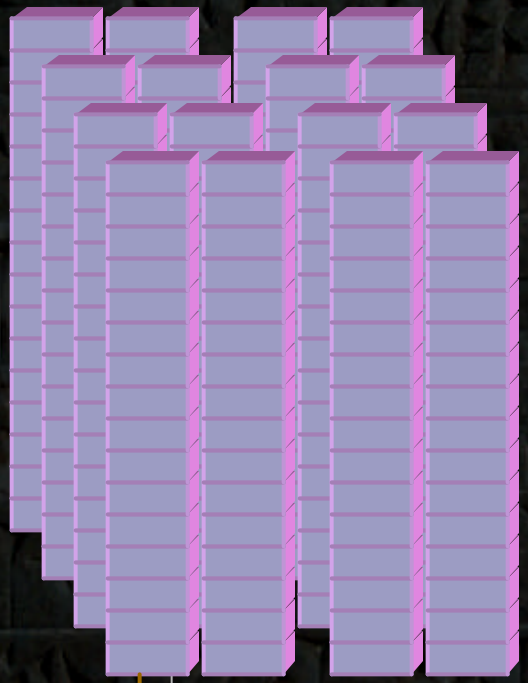


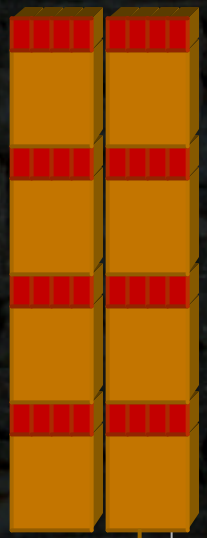
1 Visualization Town
32 Pentium III systems
with Matrox G400 cards

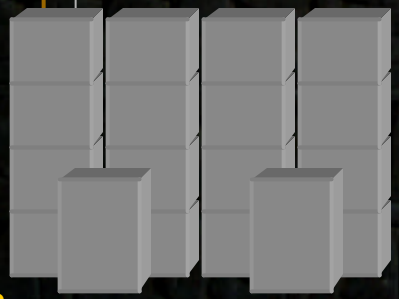
8 Computing Towns
256 Pentium III systems

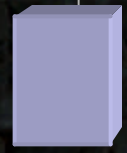
1 Storage Town
8 Xeon systems
with 300G disk each











Cluster Management
12 PIII Mayor Systems
4 PIII Front End Systems
2 Xeon File Servers
3.4 TB disk

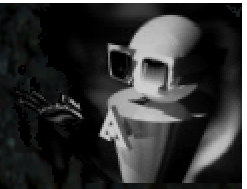
High Performance Net
64-bit Myrinet

Management Net
Gigabit and Fast Ethernet
Gigabit External Link

MCS ✦ FUTURESLAB

ARGONNE ✦ CHICAGO

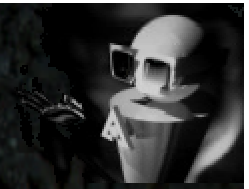
Chiba City System Details



- Purpose:
  - Scalable CS research
  - Prototype application support
- System - 314 computers:
  - 256 computing nodes, PIII 550MHz, 512M, 9G local disk
  - 32 visualization nodes, PIII 550MHz, 512M, Matrox G200
  - 8 storage nodes, 500 MHz Xeon, 512M, 300GB disk: 2.4TB total
  - 10 town mayors, 1 city mayor, other management systems: PIII 550 MHz, 512M, 3TB disk
- Communications:
  - 64-bit Myrinet computing net
  - Switched fast/gigabit ethernet management net
  - Serial control network
- Software Environment:
  - Linux (based on RH 6.0), plus “install your own” OS support
  - Compilers: GNU g++, etc
  - Libraries and Tools: PETSc, MPICH, Globus, ROMIO, SUMMA3d, Jumpshot, Visualization, PVFS, HPSS, ADSM, PBS + Maui Scheduler

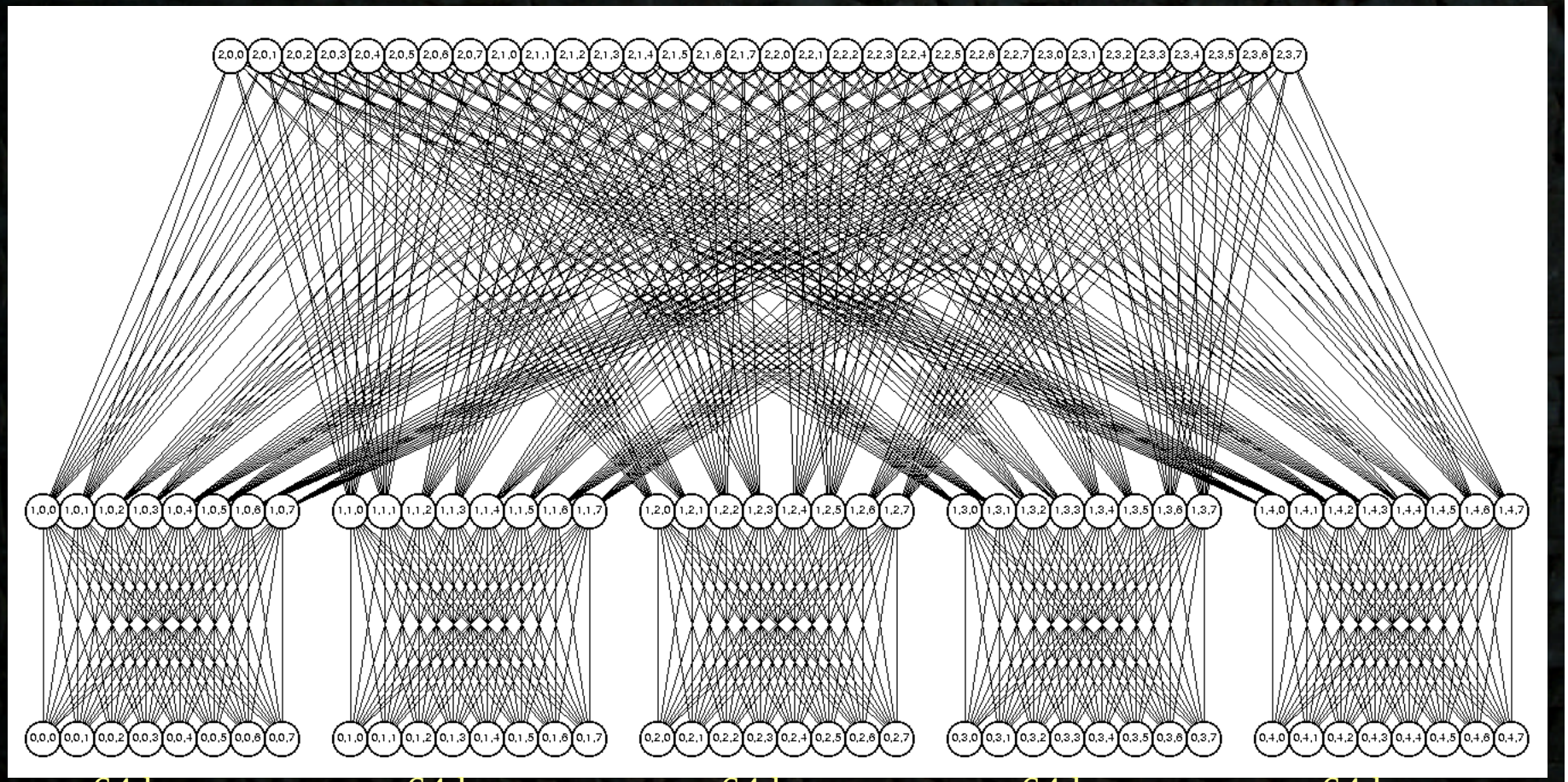

VA LINUX
Systems

Some Node Interconnect Options



- Myricom's Myrinet-2000 (~200-250MB/s)
- Compaq's ServerNet II (Beta, ~160-200+ MB/s)
- Gigabit Ethernet (~1000 Mb/s)
- Fast Ethernet (~100 Mb/s)
- OC-12 ATM (~622 Mb/s)
- Fiber Channel (~100 MB/s)
- USB (12 Mb/s)
- Firewire (IEEE 1394 400 Mb/s)

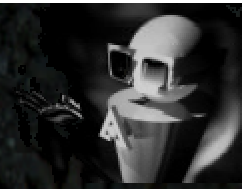
Example: 320-host Clos topology of 16-port switches



64 hosts 64 hosts 64 hosts 64 hosts 64 hosts

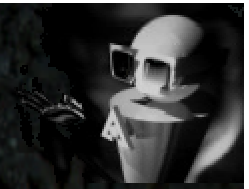
(From Myricom)

Node Issues



- Floating Point and Integer Performance
- Memory Bandwidth
- Cache Performance and Cache Sizes
- Motherboard Chip Sets

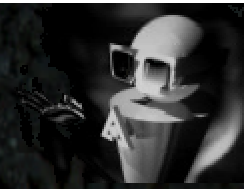
Processor Comparisons



	Alpha 21264	AMD Athlon	Intel PIII Xeon	MIPS R12000	HP PA-8500	IBM Power3	PowerPC 7400 (G4)	Sun Ultra-2	Sun Ultra-2i	Hal Sparc64-III
Clock rate	700 MHz	750 MHz	733 MHz	300 MHz	440 MHz	222 MHz	450 MHz	450 MHz	360 MHz	296 MHz
Cache (I/D)	64K/64K	64K/64K	16/16/256	32K/32K	512K/1M	32K/64K	32K/32K	16K/16K	16K/16K	64K/64K
Issue rate	4 issue	3 x86 instr	3 x86 instr	4 issue	4 issue	4 issue	3 issue	4 issue	4 issue	4 issue
Pipe stages‡	7/9 stages	9/11 stages	12/14	6 stages	7/9 stages	7/8 stages	4/5 stages	6/9 stages	6/9 stages	8/10
Out of order	80 instr	72 ROPs	40 ROPs	48 instr	56 instr	32 instr	5 instr	None	None	63 instr
Rename regs	48/41	36/36	40 total	32/32	56 total	16 int/24 fp	6 int/6 fp	None	None	34/32
BHT entries	4K×9-bit	4K×2-bit	≥512	2K×2-bit	2K×2-bit	2K×2-bit	512×2-bit	512×2-bit	512×2-bit	8K×2-bit
TLB entries	128/128	280/288	32 I/64 D	64 unified	120 unified	128/128	128/128	64 I/64 D	64 I/64 D	32/32/256
Memory B/W	2.66 GB/s	1.6 GB/s	1.06 GB/s	539 MB/s	1.54 GB/s	1.6 GB/s	1.6 GB/s	1.9 GB/s	600 MB/s	1.33 GB/s
Package	CPGA-588	CBGA-576	PGA-370	CPGA-527	LGA-544	SCC-1,088	CBGA-360	CLGA-787	PBGA-587	CLGA-957
IC process	0.25µ 6M	0.25µ 6M	0.18µ 6M	0.25µ 4M	0.25µ 4M	0.25µ 5M	0.22µ 6M	0.29µ 4M	0.29µ 4M	0.25µ 5M
Die size	205 mm²	184 mm²	106 mm²	204 mm²	477 mm²	270 mm²	83 mm²	126 mm²	150 mm²	240 mm²
Transistors	15.2 million	22 million	24 million	7.2 million	130 million	15 million	10.5 million	3.8 million	4.1 million	17.6 million
Est mfg cost*	$160	$105§	$40	$140	$330	$320	$45	$70	$85	$250
Power (max)	75 W	58 W*	24 W	20 W	50 W*	46 W	13 W	20 W	38 W	50 W
SPEC95bt†	35/55	32/24	36/31	18/30	31/49	13/28	21/20	16/24	12/17	15/28
Availability	3Q99	4Q99	4Q99	2Q99	1Q99	3Q98	3Q99	4Q98	4Q98	4Q98
1K list price	$2,296§§	$849§	$826	Not public	Not public	Not public	$345	$4,249§§	$470	Not public

†SPEC95 baseline (int/FP) ‡integer ALU/load §includes 512K L2 cache §§includes 2M L2 cache (Source: vendors, except *MDR estimates)

Benchmark Results (SPEC95baseline)



Processor	Intel PIII Xeon	Alpha 21264	AMD Athlon	HP PA-8500	IBM Pulsar	MIPS R12000	Sun Ultra-2	Hal Sparc64	PowerPC 604e	IBM Power3
System	Compaq SP750	AlphaServ. GS140-6	Microstar MS-6167†	HP9000 N4000	Bull EPC2400	SGI Origin2000	Ultra 60 Mod. 1450	Fujitsu GP7000F	RS/6000 43P-150	RS/6000 9076-N80
Clock rate	733 MHz	700 MHz	700 MHz	440 MHz	450 MHz	300 MHz	450 MHz	296 MHz	375 MHz	222 MHz
Ext cache	256K	8M	512K	none	8M	8M	4M	8M	1M	4M
099.go	36.8	33.7	32.6	34.0	29.3	16.7	18.0	15.0	16.4	15.3
124.m88Ksim	38.8	43.4	35.2	33.3	16.9	17.7	15.1	11.9	18.9	14.2
126.gcc	33.3	25.9	20.9	26.7	19.9	17.7	18.2	18.0	13.1	12.3
129.compress	23.1	27.9	24.7	29.1	19.8	17.8	18.6	15.3	11.1	13.6
130.li	44.2	35.4	36.5	33.0	17.2	13.6	13.1	12.8	12.7	12.0
132.ijpeg	32.4	41.9	26.1	24.9	13.1	15.8	16.6	15.4	17.1	14.2
134.perl	38.8	33.7	40.4	27.7	18.8	21.0	15.6	14.8	15.9	9.32
147.vortex	42.2	39.7	30.3	40.0	20.3	27.0	15.4	15.3	12.4	12.9
SPECint95b*	**35.6**	**34.7**	**31.7**	**30.8**	**19.0**	**18.1**	**16.2**	**14.7**	**14.5**	**12.8**
101.tomcatv	52.9	99.7	32.5	83.3	26.9§	36.7	32.9	58.5	10.8	45.2‡
102.swim	96.8	85.8	52.8	121	24.0§	50.9	44.2	67.9	17.1	50.6‡
103.su2cor	16.7	21.7	13.0	29.0	11.2§	18.0	15.8	16.0	4.76	16.9‡
104.hydro2d	17.6	51.8	12.0	23.6	12.4§	24.4	16.1	22.8	4.29	22.0‡
107.mgrid	29.4	73.2	15.0	39.7	20.2§	34.6	25.3	26.4	7.81	27.9‡
110.applu	17.6	26.2	13.4	33.4	15.4§	17.8	12.9	14.5	5.60	18.0‡
125.turb3d	26.5	30.9	17.0	40.6	23.0§	23.0	20.7	18.4	12.6	34.0‡
141.apsi	23.1	59.0	24.4	40.1	17.7§	28.0	30.4	20.7	11.7	22.3‡
145.fpppp	40.5	102	81.4	84.8	36.9§	47.9	26.8	38.9	35.6	37.3‡
146.wave5	37.0	67.7	35.7	59.3	26.5§	39.7	30.5	32.3	9.07	20.8‡
SPECfp95b*	**30.6**	**54.5**	**24.0**	**48.7**	**20.2§**	**30.1**	**23.9**	**27.7**	**9.76**	**27.6‡**

*SPEC95 baseline †motherboard §from Northstar-300 in RS/6000 H70 ‡from Power3-200 in RS/6000 43P-260 (Source: SPEC, AMD)

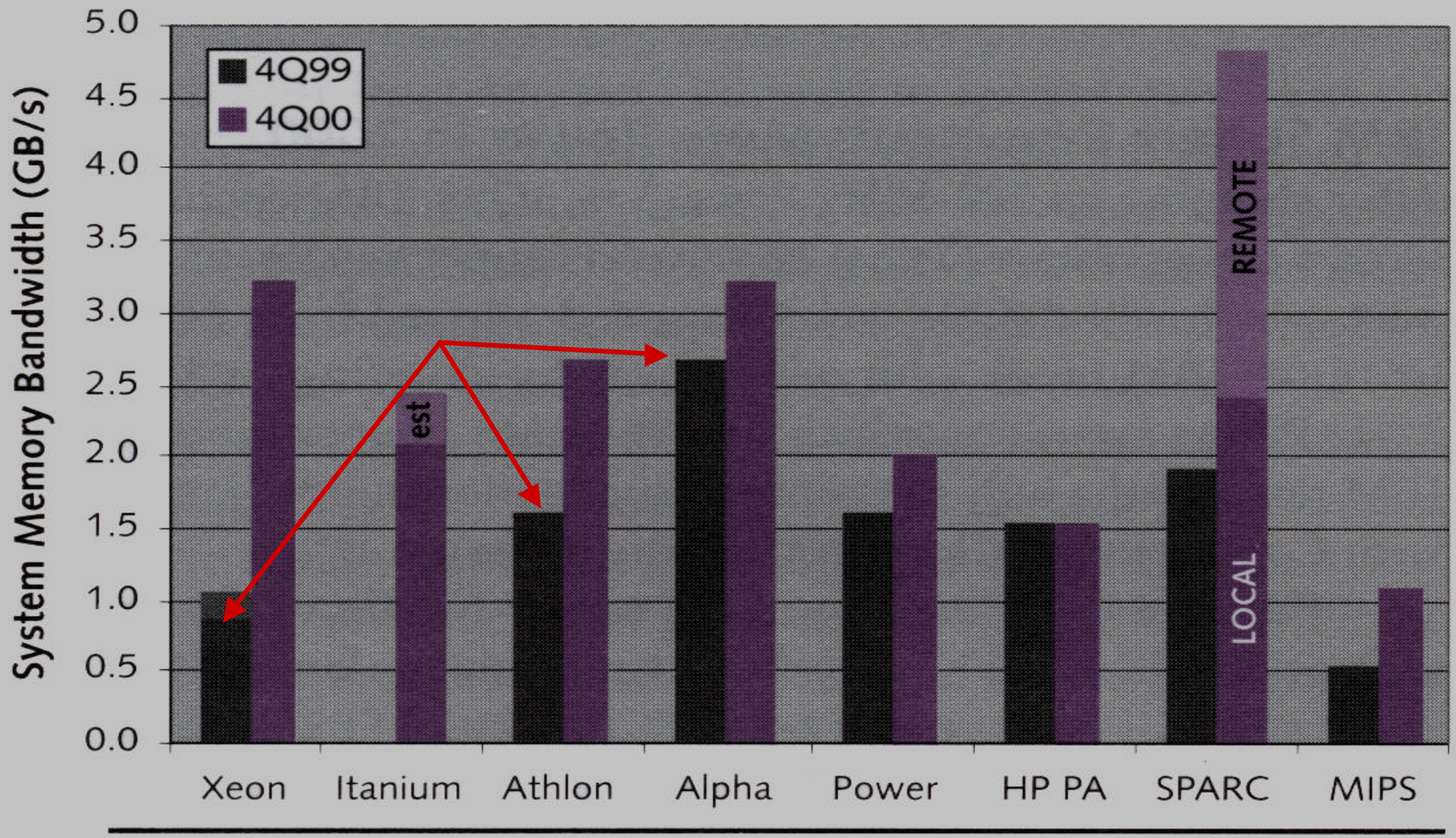


Figure 3. The Alpha 21264 has a lead today in memory (frontside bus) bandwidth, but others will close the gap next year. Sun's UltraSparc-3 looks to be the bandwidth leader, but its bandwidth is split between local and remote memory. (Source: vendors, MDR projections)

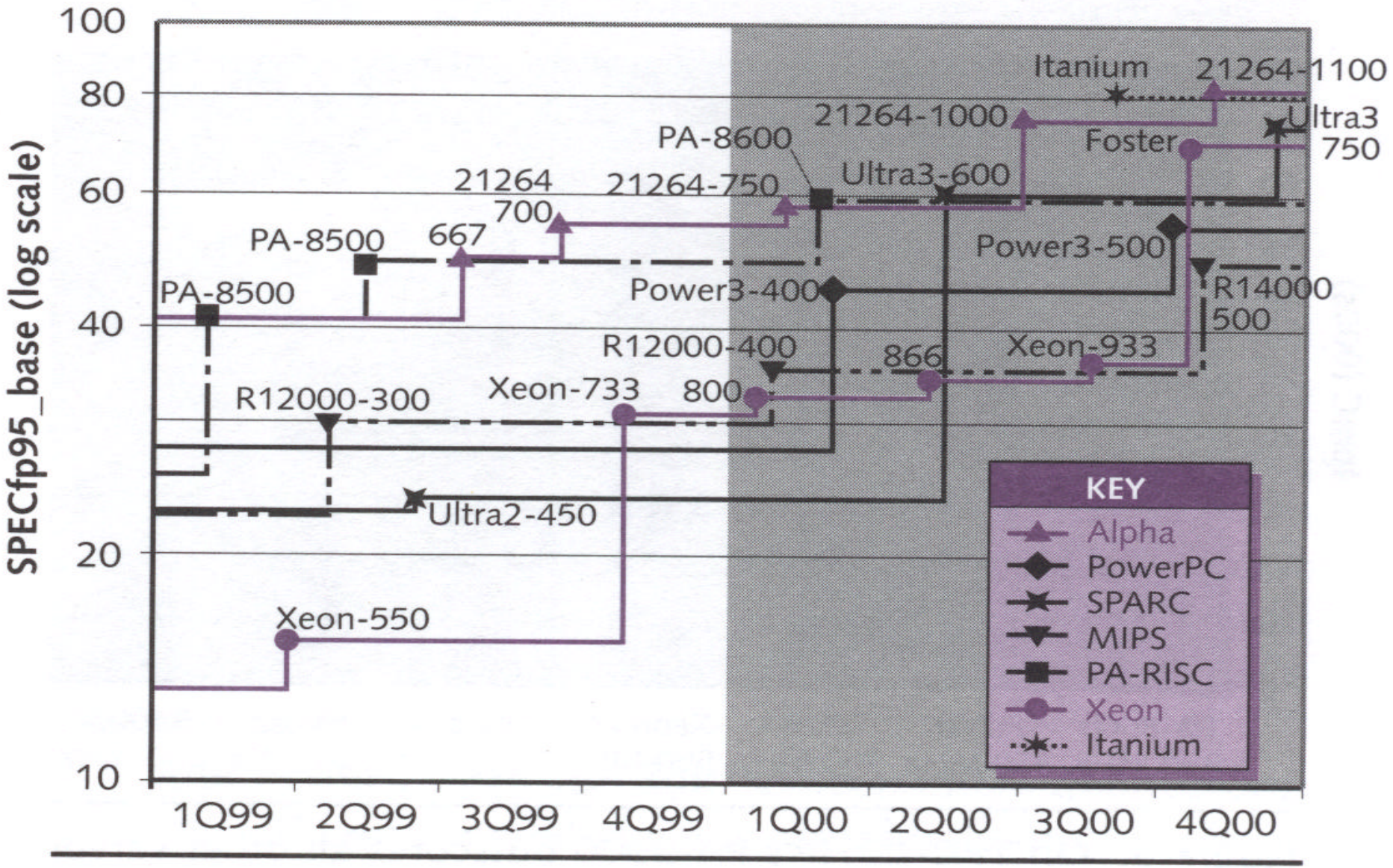
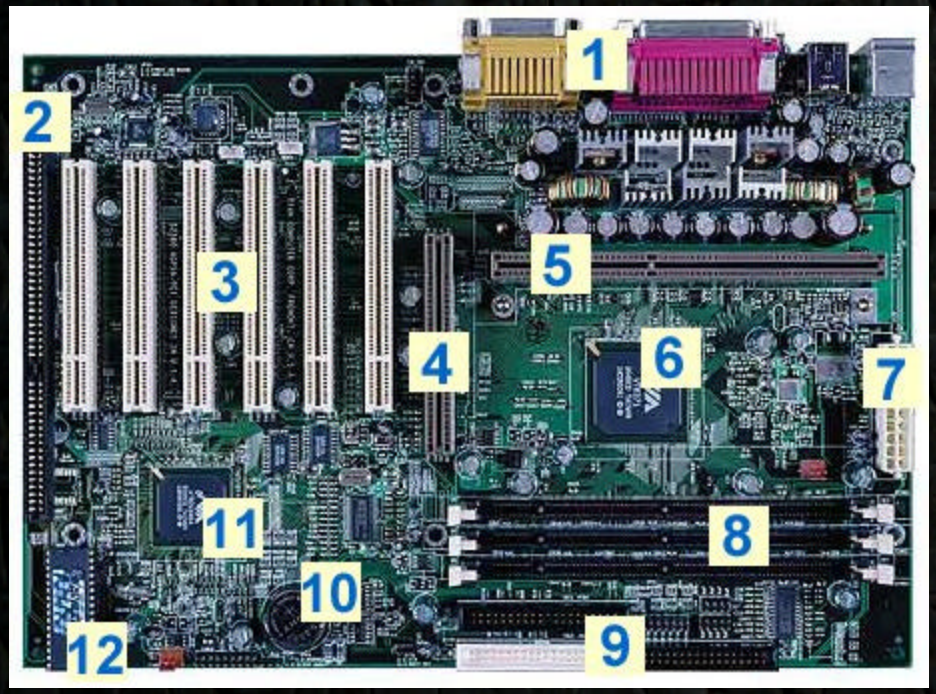


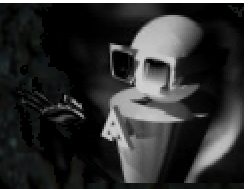
Figure 4. On SPECfp95 (base), the Alpha 21264 holds the lead but may be challenged in 2000 by HP's PA-8600 and Intel's Itanium. (Source: *www.spec.org*, MDR projections)

Motherboards and Chipsets (see www.motherboard.org)



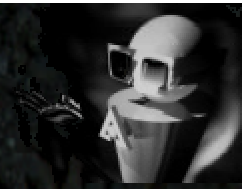
1 Ports
2 ISA Slot
3 PCI Slots
4 AGP Slot
5 CPU Slot
6 Chipset (Northbridge)
7 Power connector
8 Memory sockets
9 I/O connectors
10 Battery
11 Chipset (Southbridge)
12 BIOS chip

Your Chipset Determines



- Memory type: FPM, EDO, BEDO, SDRAM, parity-checking, ECC
- Secondary cache: burst, pipeline burst, synchronous, asynchronous
- CPU type: 486, P-24T, P5, P54C/P55C, Pentium II, Pentium III
- Maximum memory bus speed: 33, 40, 50, 60, 66, 75, 83, 100,133 MHz
- PCI bus synch: synchronous or asynchronous to memory bus speed
- PCI bus type: 32-bit or 64-bit
- SMP capability: single, dual, trio, or quad CPU support
- Support for features like: AGP, IrDA, USB, PS/2 mouse
- Support for built-in PCI EIDE controller and every possible EIDE feature you can imagine: DMA mode, PIO mode, ATA/33, etc.
- Built-in PS/2 mouse, keyboard controller and BIOS, and real-time clock circuitry

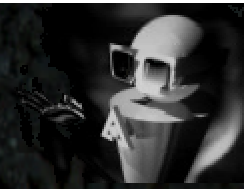
Northbridge and Southbridge



- Northbridge – In the chipset community, this refers to the major bus controller circuitry, like the memory, cache, and PCI controllers. The north bridge may have more than one discrete chip. The entire chipset is named after the numbers on the primary or largest north bridge chip. e.g. "FW82439HX" designates the Intel 430HX PCIset.

- Southbridge – In chipset lingo, this refers to the peripheral and non-essential controllers, like EIDE and serial port controllers. The south bridge usually has only one discrete chip, and has the benefit of being interchangeable on many different chipsets, for example the SiS 5513 and the Intel PIIX.

Slots, Sockets and Caches

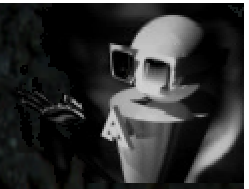


- Intel uses the P6 bus interface, in the form of Slot 1 and Socket 370. AMD's Athlon uses Slot A. * L2 cache is off chip, on the processor module.

Company	Intel			AMD				VIA-Cyrix
Product	Celeron	Pentium III		K6-2	K6-III	Athlon		MII
Code Name	Mendocino	Katmai	Coppermine		Sharptooth	K7	K75	
Process	0.25µ	0.25µ	0.18µ	0.25µ	0.25µ	0.25µ	0.18µ	0.18µ
Max. Speed	500 MHz	600 MHz	800 MHz	533 MHz	450 MHz	700 MHz	800 MHz	PR433
Interface	Socket 370	Slot 1	Socket 370	Socket 7	Socket 7	Slot A	Slot A	Socket 7
L1 Cache	32K	32K	32K	64K	64K	128K	128K	64K
L2 Cache	128K	512K*	256K	none	256K	512K*	512K*	none
Die Size	154 mm²	128 mm²	103 mm²	78 mm²	118 mm²	184 mm²	102 mm²	88 mm²
Transistors	19 million	9.5 million	23 million	9.3 million	21.3 million	22 million	22 million	6.5 million
Price Range	$64–$167	$173–$465	$284–$851	$61–$167	$163–$173	$209–$699	$799-$849	$28–$49

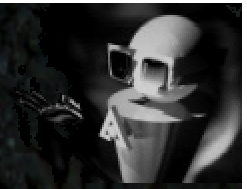
(From MDR)

Some Cluster Node Options



- Intel Pentium III
- AMD Athlon
- Compaq Alpha
- Intel IA-64
- IBM PowerPC Power4

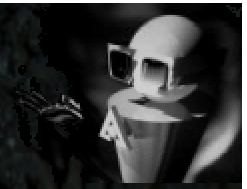
Intel Pentium III



- Now available at 1.0GHz (1000MHz) for the desktop
- 866, 850, 800, 750, 733, 700, 667, 650, 600, 550, 533, 500 and 450 MHz Clock Speeds
- 70 New Instructions (since first Pentium)
- P6 Architecture
- 133- or 100-MHz System Memory Bus
- 512K Level Two Cache or 256K Advanced Transfer Cache
- Intel's 1GHz Pentium III is shipping now at a list price of $990.

- For more information, check out Intel's Web site at www.intel.com/PentiumIII/.

IA-32 Willamette Demo'd @ 1.5 GHz



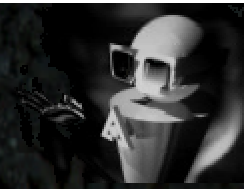
- Willamette bus is a source-synchronous 64-bit 100MHz bus that is quad-pumped delivering a total of 3.2GB/s of bandwidth--three times the bandwidth of the fastest Pentium III bus.

- A unique and unexpected aspect of Willamette's microarchitecture is its "double-pumped" ALUs. Claiming the effective performance of four ALUs, the two physical ALUs are each capable of executing an operation in every half-clock cycle.

- 20 Stage Pipeline compared to 10 on P6;

- SSE 2 includes support for (dual) double-precision SIMD floating-point operations and uses the MMX register set.

AMD Athlon v Pentium III

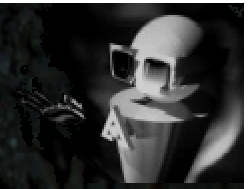


- Pentium III's smaller, but faster, on-chip L2 cache is superior to Athlon's larger, but slower, external L2 cache for most benchmarks.
- Pentium III's cache scales with core CPU frequency, while Athlon's does not. Athlon's deficiency in this regard will not be remedied until AMD delivers Thunderbird, which will have a 256K on-chip L2.

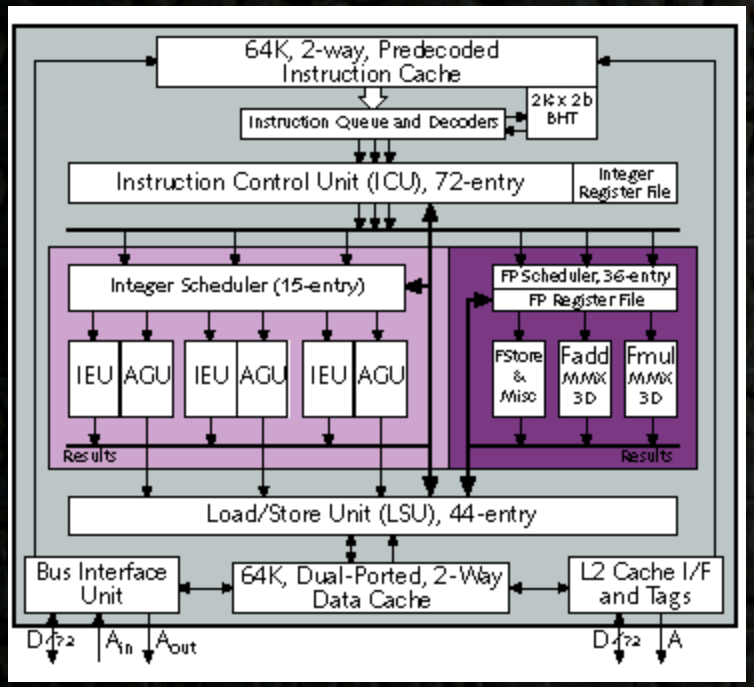
	K7		Katmai	
ROP Issue Rate	Multiply	Add	Multiply	Add
Floating Point	1 instr	1 instr	0.5 mul or 1 add	
MMX	1 instr	1 instr	1 instr	1 instr
3DNow/KNI	1 instr	1 instr	0.5 instr	0.5 instr
SIMD Width	2 ops	2 ops	4 ops	4 ops
Cycles	**Latency**	**Throughput**	**Latency**	**Throughput**
FP +	4 cycles	1 cycle	3 cycles	1 cycle
FP ×	4 cycles	1 cycle	5 cycles	2 cycles
MMX +	2 cycles	1 cycle	1 cycle	1 cycle
MMX ×+	3 cycles	1 cycle	3 cycles	1 cycle
3DNow/KNI +	4 cycles	1 cycle	4 cycles	2 cycles
3DNow/KNI ×	4 cycles	1 cycle	6 cycles	2 cycles

(Source: AMD, Intel, MDR)

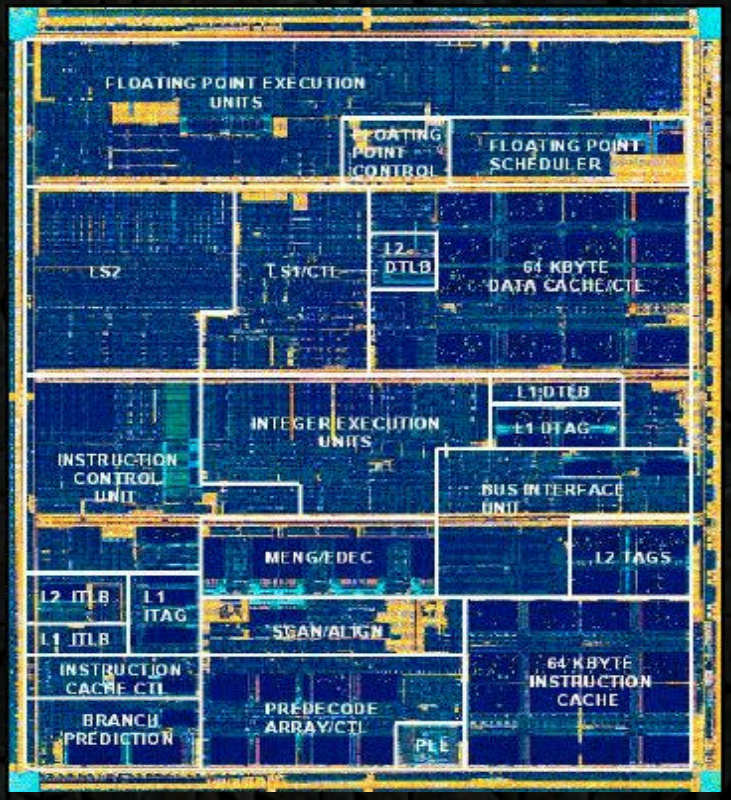
AMD K7/Athlon (www.amd.com)



- AMD is shipping 1GHz Athlon processors. List price for the 1GHz version is ~$1,000 in quantities of 1,000 units.

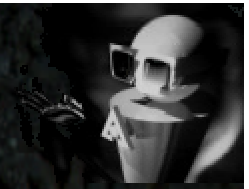


Up to 72 instructions can be in execution in K7's out-of-order integer pipe (light purple), floating-point pipe (dark purple), and load/store unit. (Source AMD and MDR)

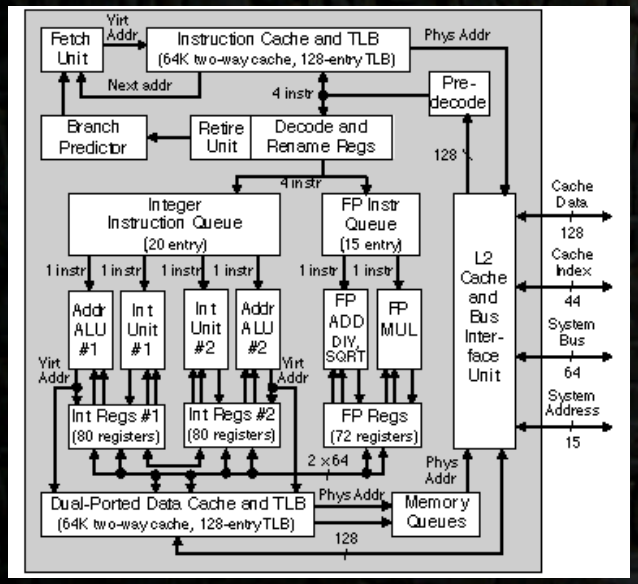


K7's 22 million transistors occupy 184 mm2 in a slightly enhanced version of AMD's 0.25-micron 6-layer-metal CS44E process. (Source: AMD)

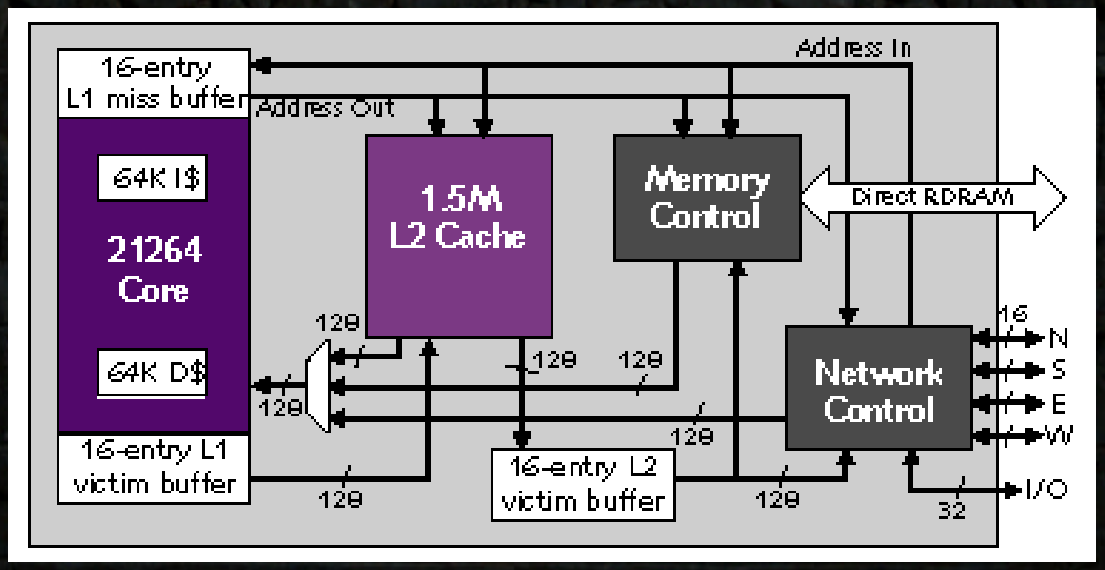
Compaq Alpha 21264/21364



- The 733-MHz four-issue out-of-order 21264 is today's fastest microprocessor
- The 21364, code-named EV7, will use multiple Rambus channels to greatly reducing memory latency.
- The 21364 will use the 21264 core but increase frequency to 1 GHz with a 0.18-micron process, add a 1.5M on-chip L2, a 6-GByte/s memory port, and 13 GBytes/s of chip-to-chip bandwidth.

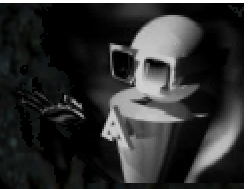


(Source Compaq, MDR)



(Source Compaq, MDR)

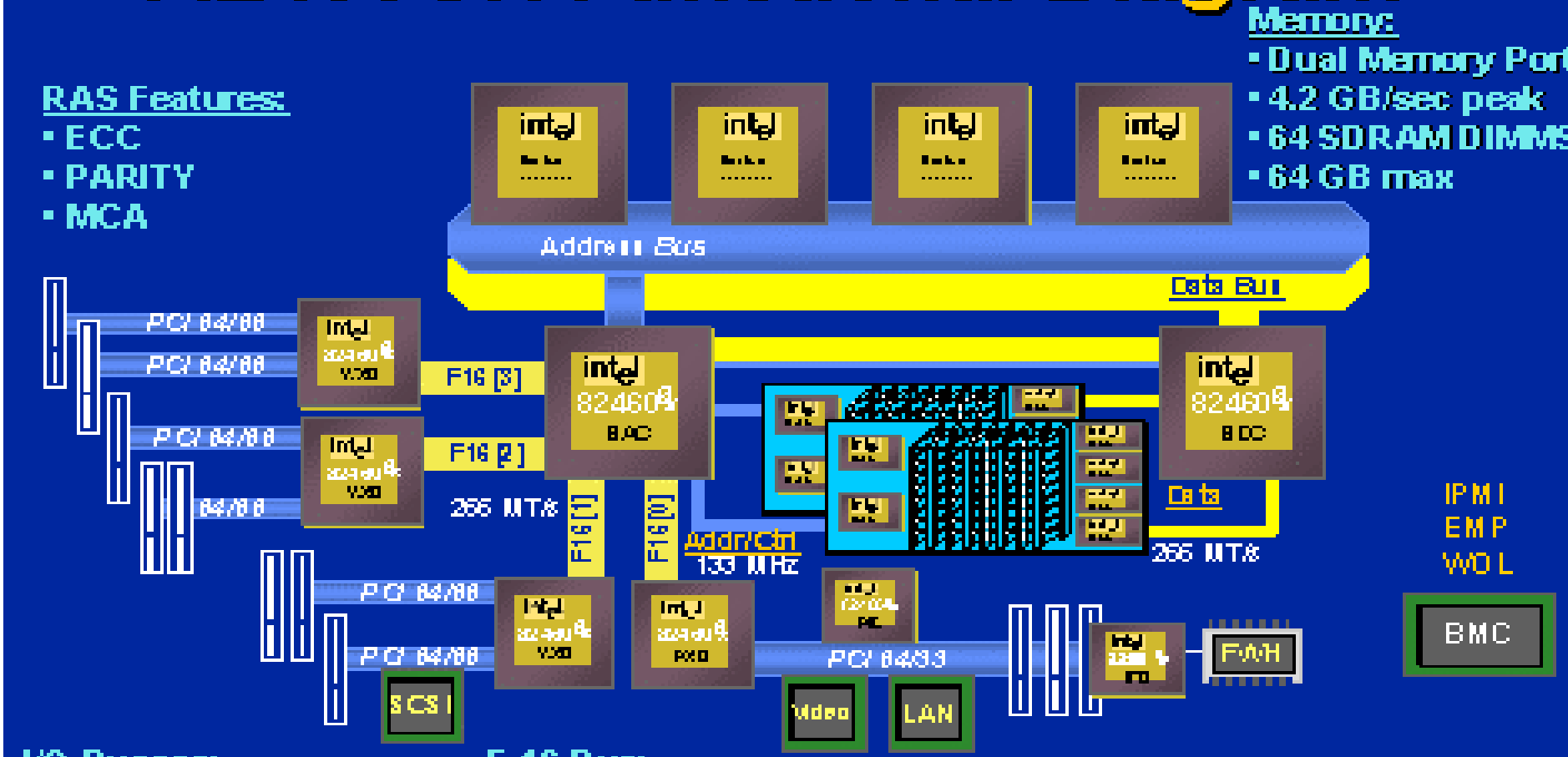
Intel IA-64



- NEW Intel/HP 64-bit microprocessor
- New Architecture Influenced by PA-RISC
  - RISC, VLIW, SuperPipelined, Superscalar
  - Executes both IA-32 and IA-64 code
- 3 Instructions (41 bit) per bundle
  - moderate VLIW, 128 bits/bundle
- Multiple Instruction bundle issues/clock (20 instructions in flight)
  - Pipeline depth of 10
- Advanced and Speculative Mechanisms
  - Memory Hierarchy Hints
  - Deep Branch Prediction

Intel IA-64 Building Blocks

AL460GX Functional Diagram



RAS Features:
- ECC
- PARITY
- MCA

Memory:
- Dual Memory Ports
- 4.2 GB/sec peak
- 64 SDRAM DIMMS
- 64 GB max

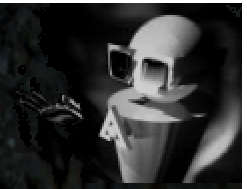
I/O Busses:
- PXB: 64b/33MHz PCI
- WXB: 64b/66MHz PHP

F-16 Bus:
- 4 pt-to-pt connects
- 16 bits wide
- 533 MB/sec each

FWH:
- 4 MB
- Block Locking

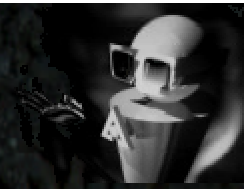
Intel Labs

Future IBM PowerPC (Power4)

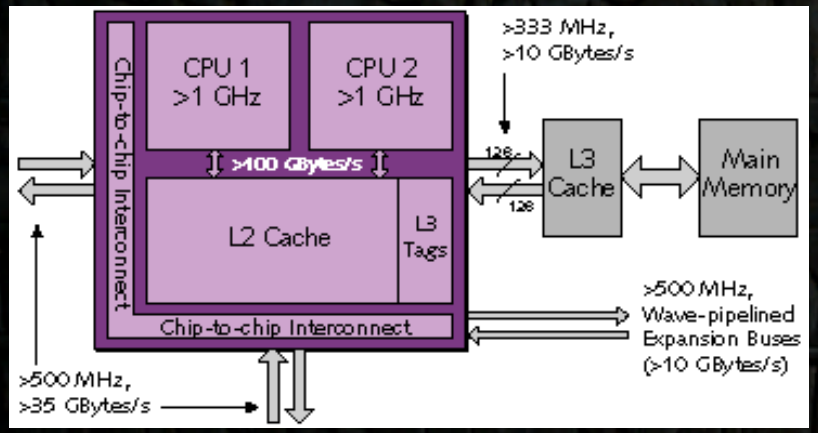


- 170-million-transistors
- Two 64-bit 1-GHz five-issue superscalar cores
- Triple-level cache hierarchy
- A 10-GByte/s main-memory interface
- 45-GByte/s multiprocessor interface.
- IBM will see first silicon on Power4 in 1Q00, and systems will begin shipping in 2H01.

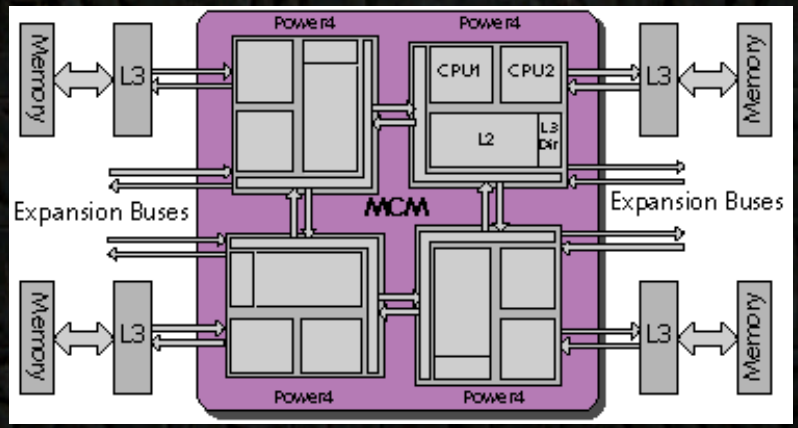
IBM's Power4



- Power4 includes two >1-GHz superscalar cores
  - > 100 GBytes/s of bandwidth to a large shared-L2 cache
  - > 55 GBytes/s of bandwidth to memory and other Power4 chips
- Four Power4 chips packaged in a MCM as an eight-processor SMP with total bandwidths of 40 GBytes/s to memory and 40 GBytes/s to other modules.



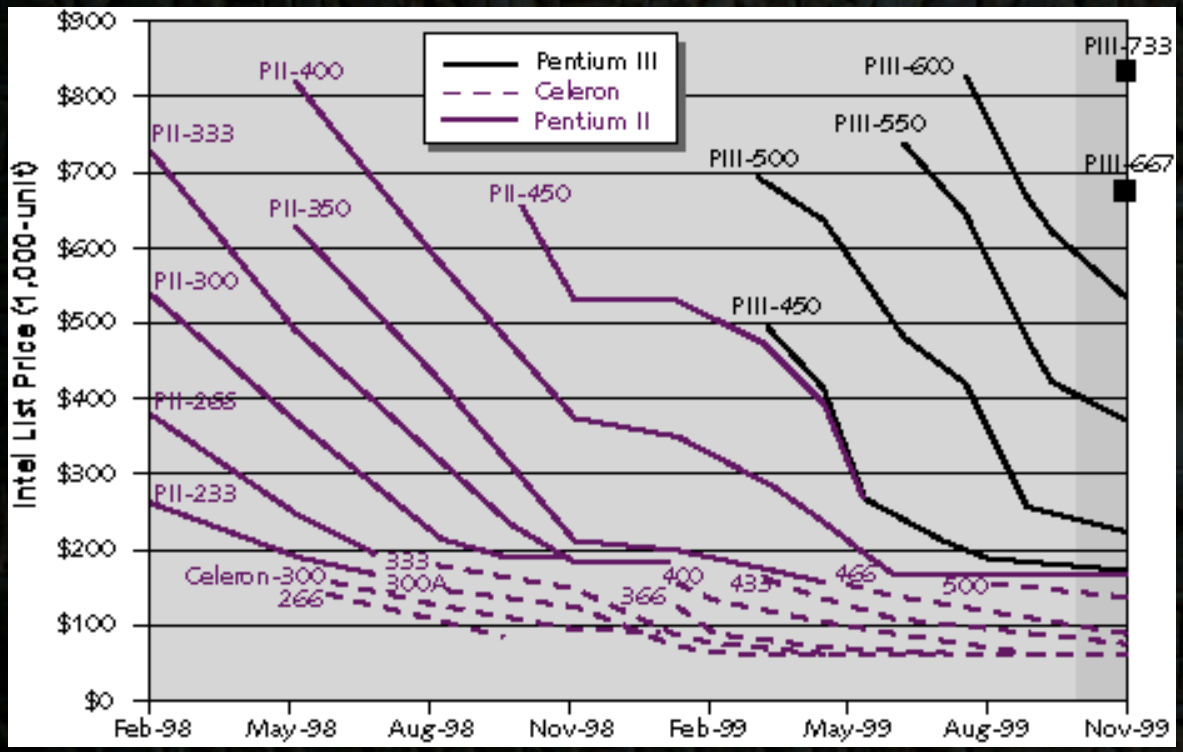
(Source IBM, MDR)



(Source IBM, MDR)

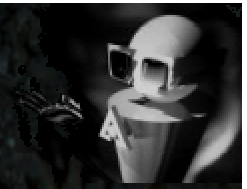
General Observations

- Frequency is likely to continue its 60% compound annual growth
- Vendors will make the next leg to two gigahertz in about 18 months, nearly 30 years faster than the one-gigahertz leg.
- Historical Intel list pricing back to 1Q98 and MDR's projected Intel pricing through 4Q99.



(Source: MDR)

Conclusions



- What is the best Node for your cluster?
- Well it depends on your application of course
  - For Integer codes both Pentium III and Athlon @ 1 GHz+
  - For 32 bit floating point IA-32 probably best price performance today
  - For 64 bit floating point Compaq Alpha is the clear lead
- How long will these choices be clear ?
  - Probably Not very long
  - Intel will introduce IA-64 (Itanium) very shortly (mid-summer)
    - ~3 GFLOPS @ 800 MHz for 64 Bit, ~6 GFLOPS for 32 bit
    - Limited by Front Side Bus bandwidth of ~ 2.1 GB/s
  - AMD's Thunderbird will address external cache issues with Athlon
  - AMD is targeting Sledgehammer at IA-64
  - Compaq is revising Alpha (EV6-> EV7) to compete with IA-64
  - IBM is putting considerable resources into Power4, and CMP and MCM's will enable IBM to produce very powerful servers/cluster nodes